New Developments in ML-based Drug Discovery

The Latest from the Conference on Neural Information Processing Systems 2020

Gabriel Ong[1]

[1]Karanicolas Lab
Program in Molecular Therapeutics - Fox Chase Cancer Center

Group Meeting, 13th January 2021

Table of Contents

Table of Contents

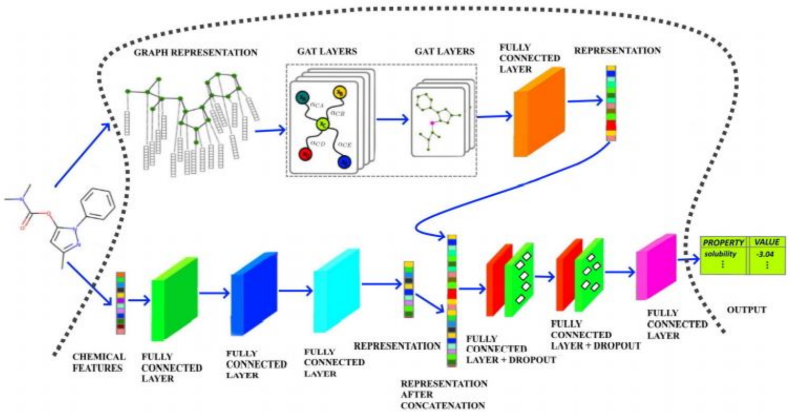
1. DNNs to Predict Properties of Drug-Like Molecules
2. Transformer-Based Molecule Encoding
3. Geometric Vector Perceptrons for Protein Structure
4. Conclusions and Thanks

Deep Neural Network Approach to Predict Properties of Drugs and Drug-Like Molecules (Wiercioch and Kirchmair)

Designing a deep neural network to predict molecular properties (solubility, lipophilicity, etc.)

- Utilizes 2 separate blocks of representations, one feature-based and one graph-based.
- Outperforms state-of-the-art models on MoeculeNet, a standard benchmark for molecular ML.

Model Architecture

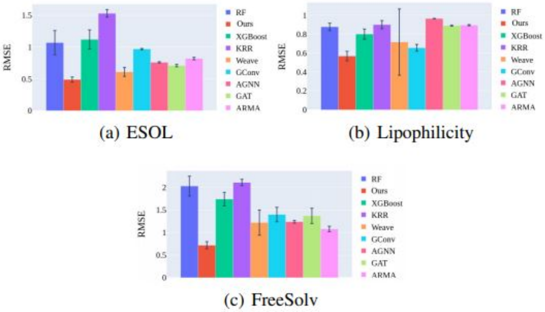


Methods

Model is trained with data randomly split 80-10-10. Final results are averaged from an ensemble of 10 models.

- Classification tasks are evaluated by AUC-ROC.
- Regression tasks are evaluated by RMSE.

Results (Regression)



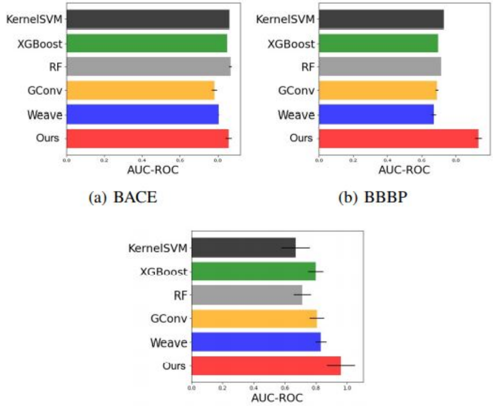
(a) ESOL

(b) Lipophilicity

(c) FreeSolv

Figure 2: The RMSE scores of various methods on regression task and test set. We achieved the least RMSE (lower is better).

Results (Classification)



(a) BACE

(b) BBBP

(c) ClinTox

Figure 3: The AUC-ROC scores of various methods on classification task and test set. We achieved higher AUC-ROC score.

Table of Contents

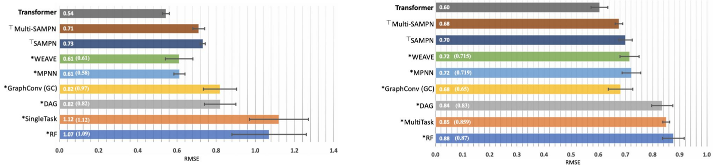
1. DNNs to Predict Properties of Drug-Like Molecules
2. Transformer-Based Molecule Encoding
3. Geometric Vector Perceptrons for Protein Structure
4. Conclusions and Thanks

Transformer-Based Molecule Encoding (Nayak, *et. al.*)

Searching for better ways to parametrize molecules and predict their properties.

- Using self-attention layers and transformer neural networks, the resulting model extracts features more efficiently on small data sets.
- Current molecular encoding techniques leverage message passing neural networks (MPNNs) that require large amounts of training data, often unavailable in drug-discovery data sets.

Results

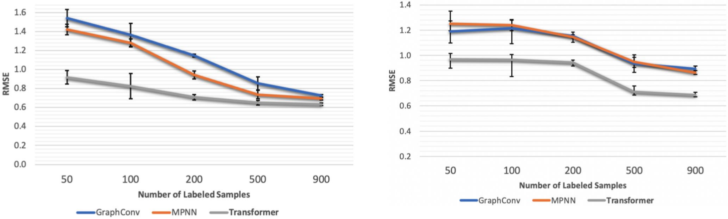


(a) ESOL

(b) Lipophilicity

Figure 1: Test RMSE results on ESOL and Lipophilicity datasets
$*$ models from Wu et al.[13]. $\top$ models from Tang et al.[11]

Results



(a) ESOL

(b) Lipophilicity

Figure 2: Label ablation experiment varying number of sample labels seen during training MPNN and GraphConv implementations are from Wu et al.[13].

Table of Contents

1. DNNs to Predict Properties of Drug-Like Molecules
2. Transformer-Based Molecule Encoding
3. Geometric Vector Perceptrons for Protein Structure
4. Conclusions and Thanks

Learning from Protein Structure from Geometric Vector Perceptrons (Jing, *et.al.*)

Better ways to learn from protein structure.

- Geometric vector perceptrons allow neural networks to understand both geometric and relational aspects of 3D biomolecular structure.
- Geometric vector perceptrons replace dense layers in a graph neural network operating on collections of vectors.
- Approach demonstrated on protein design task.

What is a Geometric Vector Perceptron?

- Previous architectures have typically encoded 3D structures by encoding vector features (node orientation and edge direction) as scalars in a local coordinate system.
- Geometric vector perceptrons allow these features to be encoded as geometric vectors in a global coordinate system at all levels of the neural network.

Results (Synthetic Task)

Table 1: Comparison of the CNN, standard GNN, and GVP-GNN on the three objectives on the synthetic test set. The MSE losses are standardized such that predicting a constant value (i.e. the mean) would result in unit loss. Results are reported as the mean ± S.D. over five randomly shuffled splits.

Model	Parameters	Off-center (geometric)	Perimeter (relational)	Combined
CNN	59k	0.319 ± 0.014	0.532 ± 0.028	0.522 ± 0.016
Standard GNN	40k	0.871 ± 0.045	0.128 ± 0.009	0.421 ± 0.025
GVP-GNN	22k	**0.206 ± 0.024**	**0.106 ± 0.006**	**0.155 ± 0.024**

Results (Computational Protein Design)

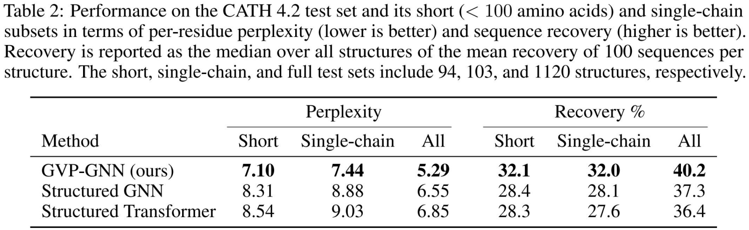


Table 2: Performance on the CATH 4.2 test set and its short (< 100 amino acids) and single-chain subsets in terms of per-residue perplexity (lower is better) and sequence recovery (higher is better). Recovery is reported as the median over all structures of the mean recovery of 100 sequences per structure. The short, single-chain, and full test sets include 94, 103, and 1120 structures, respectively.

	Perplexity			Recovery %		
Method	Short	Single-chain	All	Short	Single-chain	All
GVP-GNN (ours)	**7.10**	**7.44**	**5.29**	**32.1**	**32.0**	**40.2**
Structured GNN	8.31	8.88	6.55	28.4	28.1	37.3
Structured Transformer	8.54	9.03	6.85	28.3	27.6	36.4

Table of Contents

1. DNNs to Predict Properties of Drug-Like Molecules
2. Transformer-Based Molecule Encoding
3. Geometric Vector Perceptrons for Protein Structure
4. Conclusions and Thanks

Thanks and Acknowledgements

Thanks to:

- John Karanicolas
- Kiruba Palani
- Jake Khowsathit
- Chris Parry
- Grigorii Andrianov

FOX CHASE CANCER CENTER
TEMPLE HEALTH

- Daniel Yeggoni
- Sven Miller
- Lei Ke

References

1. Jing, *et.al.* 'Learning from Protein Structure with Geometric Vector Perceptrons'. *ML For Molecules, 2020 Conference on Neural Information Processing Systems*. December 2020.
2. Nayak, *et.al.* 'Transformer based Molecule Encoding for Property Prediction'. *ML For Molecules, 2020 Conference on Neural Information Processing Systems*. December 2020.
3. Wieroch and Kirchmair. 'Deep Neural Network Approach to Predict Properties of Drugs and Drug-Like Molecules'. *ML For Molecules, 2020 Conference on Neural Information Processing Systems*. December 2020.